TP4 : RDF et SPARQL

UCBL - Base de données pour la bioinformatique - 2024 / 2025

Objectif du TP : écrire des requêtes SPARQL sur des données RDF

1 Préparation de l'environnement

Une fois n'est pas coutume, nous utiliserons un SGBD léger et automatiquement configuré pour gérer des triplets RDF et les interroger avec SPARQL, en l'occurrence le "triple store" BlazeGraph. Si vous travaillez localement[1], une alternative possible (également open-source et multi-plateformes) est GraphDB.

Une instance de BlazeGraph tourne sur `http://192.168.77.137:8080/`. Si vous vous connectez depuis l'extérieur du campus, il faut utiliser le VPN de l'université.

BlazeGraph[2] utilise des *namespaces* pour organiser les données, un namespace pouvant être vu comme l'équivalent d'une base de données en relationnel. Dans l'interface web, réalisez les étapes suivantes :

- Cliquez sur l'onglet `Namespaces` ;
- Dans le tableau, visualisez la ligne du namespace `bdbio` et cliquez sur `Use`. Attention : le SGBD est partagé, donc ne pas utiliser les autres options.

Le jeu de données pour ce TP est un extrait d'un génome humain en RDF[3]. Les triplestores ne permettent pas de visualiser facilement les données (le graphe serait de toute façon peu exploitable). L'onglet `Explore` permet de lister les liens entrants et sortants d'une ressource, par exemple testez pour cette ressource protéine `<http://purl.uniprot.org/uniprot/P04275>`.

2 Exécution de requêtes SPARQL

Pour exécuter une requête SPARQL avec BlazeGraph, il faut la saisir dans la zone de texte de l'onglet `Query` et cliquer sur le bouton `Execute`. Vous pouvez écrire vos requêtes dans un fichier texte ou directement dans l'interface. Dans les deux cas, n'oubliez pas de sauvegarder régulièrement.

Préfixes Uniprot pour certaines requêtes :

```
PREFIX unic: <http://purl.uniprot.org/core/>
PREFIX unip: <http://purl.uniprot.org/uniprot/>
```

Requêtes à traduire en SPARQL :

1. L'ensemble des triplets (137992 résultats)
2. Liste des classes (92 résultats)
3. Les ressources de type protéines (97 résultats)
4. Les informations sur chaque protéine ou chaque gène (sous forme de triplets). Les informations seront triées par identifiant de ressource (18539 résultats)
5. Le nombre d'informations (de triplets) pour la protéine *P04275* (résultat = 661)
6. Le nom complet de chaque protéine, et si disponible, son nom court. Il faut utiliser, à partir d'une protéine, le prédicat `recommendedName` (95 résultats)

[1] Si vous utilisez un SGBD en local, il faut évidemment importer le jeu de données (fichier RDF du protéome humain).
[2] Documentation de BlazeGraph, https://github.com/blazegraph/database/wiki/Main_Page
[3] Source Uniprot #9606, `https://www.uniprot.org/taxonomy/9606`

7. Pour chaque protéine ayant au moins une interaction, donner son nombre d'interactions (56 résultats). Le prédicat `interaction` est utilisé pour lier une ressource protéine à une interaction.
8. Les ressources protéines avec leurs articles publiés après 2014 (33 résultats). Chaque article devra avoir son titre, son année et son journal de publication, et ses auteur.e.s concaténés en une seule valeur.